(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
17 January 2002 (17.01.2002)

PCT

(10) International Publication Number
WO 02/05084 A2

(51) International Patent Classification⁷: G06F 7/08

(21) International Application Number: PCT/EP01/07801

(22) International Filing Date: 6 July 2001 (06.07.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

00114636.4	7 July 2000 (07.07.2000)	EP
00115867.4	24 July 2000 (24.07.2000)	EP
00125503.3	21 November 2000 (21.11.2000)	EP

(71) Applicant (*for all designated States except US*): LION BIOSCIENCE AG [DE/DE]; Im Neuenheimer Feld 515, 69120 Heidelberg (DE).

(72) Inventor; and
(75) Inventor/Applicant (*for US only*): MINCH, Eric [US/DE]; Altes Holz 4, 69207 Sandhausen (DE).

(74) Agent: BOEHMERT & BOEHMERT; Schohe, Stefan, Hollerallee 32, 28209 Bremen (DE).

(81) Designated States (*national*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: METHOD AND APPARATUS FOR ORDERING ELECTRONIC DATA

(57) Abstract: The present invention relates to the field of management of data in a computer system. The invention proposes a new way of automatically ordering data and arranging them in a data structure in a computer. The invention employs the distance as a measure of similarity between data sets. Data sets are assigned to a structure of clusters depending on whether they have a distance above or below a limiting value that is correlated with a peak in the density of distance values.

Claims

1. Method of automatically ordering a plurality of sets of electronic data by means of a data processing unit, comprising the following steps to be performed by said data processing unit:

- at least for a selected group of data sets, determining the distance D between any two data sets, said distance being defined as a function of a pair of two data sets, rendering a numerical value, said function having a first value D_0 defined for the case of a pair of identical data sets, the difference of the distance D of any pair to said value D_0 being defined to be either greater than or equal zero for all pairs, $D-D_0 \geq 0$, or less than or equal zero for all pairs, $D-D_0 \leq 0$,
- determining the density of distance values over the range of determined distance values,
- determining one or more limiting values, at least some of the limiting values defining an upper boundary of a peak in said density of distance values, respectively, if said difference is defined to be $D-D_0 \geq 0$ for all pairs, and at least some of the limiting values defining a lower boundary of a peak, respectively, if said difference is defined to be $D-D_0 \leq 0$, said limiting values forming an increasing series in case of a plurality of limiting values,
- creating correlation data correlating each data set to a cluster in a hierarchy of clusters, the number of cluster levels in said hierarchy corresponding to the number of limiting values, wherein,

if said difference is defined to be $D-D_0 \geq 0$ for all pairs,

- the data sets contained in each first level cluster in said hierarchy are related to one another in that for each data set the minimum pairwise distance to other data sets in said cluster is less than the lowest limiting value,
- each higher order cluster in said hierarchy comprises data sets of a group of one or more clusters of lower levels, wherein, if said group comprises more than one cluster, each cluster in this group forms a pair with another cluster in this group for which pair there is at least one data set of one cluster of said pair having a distance from a data set of the other cluster of said pair which is less than that

limiting value that is the next higher one in said increasing series of limiting values to that limiting value defining clusters at the next lower level,

and, if said difference is defined to be $D\text{-}D_0 \leq 0$ for all pairs,

- the data sets contained in each first level cluster in said hierarchy are related to one another in that for each data set the maximum pairwise distance to other data sets in said cluster is greater than the highest limiting value,
- each higher order cluster in said hierarchy comprises data sets of a group of one or more clusters of lower levels, wherein, if said group comprises more than one cluster, each cluster in this group forms a pair with another cluster in this group for which pair there is at least one data set of one cluster of said pair having a distance from a data set of the other cluster of said pair, which is greater than that limiting value that is the next lower one in said increasing series of limiting values to that limiting value defining clusters at the next lower level.

2. Method according to claim 1, characterised in that it comprises the step of creating data correlating each data set to a cluster in a hierarchy of clusters, the number of cluster levels in said hierarchy corresponding to the number of limiting values,

wherein, if said difference is defined to be $D\text{-}D_0 \geq 0$ for all pairs,

- each first level cluster in said hierarchy comprises at least one data set to which all other data sets of said cluster have a distance less than the lowest limiting value,
- each second level cluster in said hierarchy comprises at least one data set to which all other data sets of said cluster have a distance less than the second lowest limiting value,
- each higher order cluster in said hierarchy comprises at least one data set to which all other data sets of said cluster have a distance which is less than that limiting value that is the next higher one in said increasing series of limiting values to that limiting value defining clusters at the next lower level,

and, if said difference $D\text{-}D_0 \leq 0$ for all pairs,

- each first level cluster in said hierarchy comprises at least one data set to which all other data sets of said cluster have a distance greater than the highest limiting value,

- each second level cluster in said hierarchy comprises at least one data set to which all other data sets of said cluster have a distance greater than the second highest limiting value,
- each higher order cluster in said hierarchy comprises at least one data set to which all other data sets of said cluster have a distance which is greater than that limiting value that is the next lower one in said increasing series of limiting values to that limiting value defining clusters at the next lower level.

3. Method according to claim 1 or 2, characterized in that said data correlating the data sets to clusters comprise:
   - data correlating each data set to one or more first level clusters,
   - data correlating each cluster at a level less than the highest level to a cluster or a plurality of clusters at a higher level.

4. Method according to one of claims 1 to 3, characterized by the step of controlling a display device on the basis of said correlation data to create a graphic symbolic display of clusters at one or more levels.

5. Method according to one of claims 1 to 4, characterized by the step of creating a directory structure on the basis of said correlation data, each cluster corresponding to a directory and each cluster level to a directory level.

6. Method according to one of claims 1 to 5, characterized by the step of creating a database from said data sets and said correlation data, the data model of said data base being defined by said hierarchy of clusters.

7. Method according to claim 6, characterized in that the database is a relational data base, wherein the keys are defined by cluster names and the values are defined by the name of the parent cluster.

8. Method according-to claim 6, characterized in that the database is an object oriented data base, wherein the keys are defined by cluster names and the values are defined by the name of the parent directory.

9. Method according to one of claims 1 to 8, characterized in that a group of data sets comprising one or more predetermined data elements is selected and said limiting values are determined on the basis of said selected group of data sets.

10. Method according to one of claims 1 to 9, characterized in that the total range of distance values is completely partitioned into a sequence of distance intervals and said density of distance values is determined as the number or normalized number of distance values in each distance interval.

11. Method according to claim 10, characterized in that a plurality of partitionings of said total range of distances with increasing interval size is established and said density is established for each of said partitionings, and that preliminary limiting values are determined for each partitioning and optimized limiting values are obtained by averaging or fitting said preliminary limiting values, wherein said correlation data are established on the basis of said optimized limiting values.

12. Method according to claim 10 or 11, characterized in that a distribution of distance density values is established from said partitioning and said limiting values are determined from said distribution.

13. Method according to one of claims 1 to 12, characterized in that one or more limiting values are determined as a minimum or zero point of the density adjacent to a maximum of said density.

14. Method according to claim 1 to 13, characterized in that a curve is fitted to density values and one or more limiting values are determined as the point of a minimum or zero adjacent to a maximum of said curve.

15. Method according to claim 14, characterized in that said curve is fitted to a distribution of density values.

16. Method according to one of claims 14 or 15, characterized in that said curve is a polynomial or a trigonometric function or a function of trigonometric functions.

17. Method according to one of claims 1 to 16, characterized in that said data sets comprise text data and said distance is a function of the number of common words of two data sets.

18. Method according to one of claims 1 to 16, characterized in that said data sets comprise genetic information and said distance is a function of the number of identical data elements succeeding one another in two partial sequences in said data sets.

19. Method according to one of claims 1 to 18, characterized in that the step of creating correlation data comprises
   - establishing a distance matrix for all data sets,
   - assigning data sets to a first level cluster that are linked by matrix elements having a value less than the lowest limiting value for $D \geq D_0$ or greater than the highest limiting value for $D \leq D_0$.

20. Method according to one of claims 1 to 19, characterized in that the data sets are displayed graphically as vertices connected to every other vertex by edges, the length of each edge corresponding to the distance between two data sets, that edges having a length less than that corresponding to the lowest limiting value are removed and data sets represented by a connected remaining subgraph are assigned to the same cluster at the lowest level.

21. Database obtainable by a method of one of claims 1 to 20.

22. Computer program adapted to perform all steps of a method according to claim 1 or any claim dependent thereon.

23. Computer program according to claim 22 embodied in a computer readable medium.

24. Apparatus for automatically ordering a plurality of sets of electronic data according to their similarities, comprising data processing means performing the steps of a method according to one of claims 1 to 20.

25. Apparatus according to claim 24, characterized in that it comprises a display device and said data processing means controls said display device according to a method according to claim 4.

26. Apparatus according to one of claims 24 or 25, characterized in that it comprises data storage means for storing said data sets in a directory structure, said directory structure being obtainable according to a method according to claim 5.

27. Method of operating an apparatus for searching and/or ordering data sets, said apparatus containing or being capable of obtaining correlation data obtainable according to one of claims 1 to 20, characterized by the following steps:

- inputting data elements,
- selecting data sets comprising these data elements,
- selecting a cluster at the lowest level in a hierarchy of the selected data sets defined by said correlation data,

whereupon said apparatus outputs data related to the elements of said selected cluster.

28. Method according to claim 27, characterized in that the apparatus, having outputted data related to the elements of said selected cluster outputs data related to the elements of the next higher order cluster comprising said selected cluster and not contained in said selected cluster.

29. Method according to one of claims 27 or 28, characterized in that said apparatus proceeds with outputting data related to elements of at least one higher order cluster only upon a related input by a user.